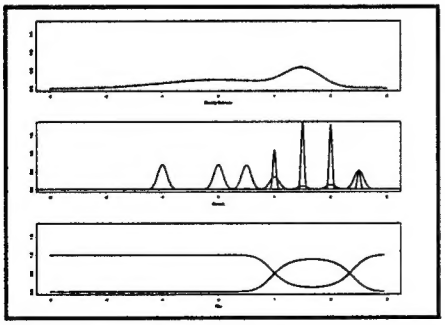
Filtered Kernel Density Estimation

David J. Marchette, Carey E. Priebe,
George W. Rogers and Jeffrey L. Solka

Technical Report No. 104
October, 1994

Center for Computational Statistics



George Mason University
Fairfax, VA 22030

CENTER FOR COMPUTATIONAL STATISTICS
TECHNICAL REPORT SERIES (RECENT REPORTS)

TR 93. Winston C. Chow, Modeling and Estimation with Fractional Brownian Motion and Fractional Gaussian Noise (Ph.D. Dissertation), February, 1994.

TR 94. Mark C. Sullivan and Edward J. Wegman, Correlation Estimators Based on Simple Nonlinear Transformations, February, 1994, To appear **IEEE Transactions on Signal Processing.**

TR 95. Mark C. Sullivan and Edward J. Wegman, Normalized Correlation Estimators Based on Simple Nonlinear Transformations, March, 1994.

TR 96. Kathleen Perez-Lopez and Arun Sood, Comparison of Subband Features for Automatic Indexing of Scientific Image Databases, March, 1994.

TR 97. Wendy L. Poston and Jeffrey L. Solka, A Parallel Method to Maximize the Fisher Information Matrix, June, 1994.

TR 98. Edward J. Wegman and Charles A. Jones, Simulating a Multi-target Acoustic Array on the Intel Paragon, June, 1994.

TR 99. Barnabas Takacs, Edward J. Wegman and Harry Wechsler, Parallel Simulation of an Active Vision Model, June, 1994.

TR 100. Edward J. Wegman and Qiang Luo, Visualizing Densities, October, 1994.

TR 101. Daniel B. Carr, Converting Tables to Plots, October, 1994.

TR 102. Julia Corbin Fauntleroy and Edward J. Wegman, Parallelizing Locally-Weighted Regression, October, 1994.

TR 103. Daniel B. Carr, Color Perception, the Importance of Gray and Residuals on a Choropleth Map, October, 1994.

TR 104. David J. Marchette, Carey E. Priebe, George W. Rogers and Jeffrey L. Solka, Filtered Kernel Density Estimation, October, 1994.

TR 105. Jeffrey L. Solka, Edward J. Wegman, Carey E. Priebe, Wendy L. Poston and George W. Rogers, A Method to Determine the Structure of an Unknown Mixture Using the Akaike Information Criterion and the Bootstrap, October, 1994.

TR 106. Wendy L. Poston, Edward J. Wegman, Carey E. Priebe and Jeffrey L. Solka, A Contribution to the Theory of Robust Estimation of Multivariate Location and Shape: EID, October, 1994.

TR 107. Clifton D. Sutton, Tree Structured Density Estimation, October, 1994.

TR 108. Charles A. Jones, Simulating a Multi-target Acoustic Array on the Intel Paragon (M.S. Thesis), October, 1994.

Filtered Kernel Density Estimation

David J. Marchette[1,3], Carey E. Priebe[2,3], George W. Rogers[1,3], Jeffrey L. Solka[1,3]

[1]Naval Surface Warfare Center, Dahlgren Div, B10
Dahlgren, Virginia 22448

[2]The Johns Hopkins University
Baltimore, Maryland 21218

[3]George Mason University
Center for Computational Statistics
Fairfax, Virginia 22030

Abstract

A modification of the kernel estimator for density estimation is proposed which allows the incorporation of local information about the smoothness of the density. The estimator uses a small set of local bandwidths rather than a single global one as in the standard kernel estimator. It uses a set of filtering functions which determine the extent of influence of the local bandwidths. Various versions of the idea are discussed. The estimator is shown to be consistent and is illustrated by comparison to the single bandwidth kernel estimator for the case in which the filter functions are derived from finite mixture models.

1. INTRODUCTION

The kernel density estimator has been studied widely since its introduction in Rosenblatt 1956 and Parzen 1962. Given i.i.d. data $x_1,...,x_n$ drawn from the unknown density α, the standard kernel estimator is the single bandwidth estimator:

$$\hat{\alpha}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right). \tag{1}$$

Much work has been done on selecting the optimal bandwidth h under different assumptions on α or different optimality criteria (see the recent books by Silverman 1986 and Scott 1992, and the bibliographies contained therein, for a good introduction to kernel estimators and bandwidth selection). Alternatively, variable bandwidth kernel estimators are of the form

$$\hat{\alpha}(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h_i} K\left(\frac{x-x_i}{h_i}\right), \tag{2}$$

or variations on this theme. One then requires a choice of many bandwidths, and several approaches have been investigated. The obvious problem which may arise in these variable bandwidth estimators is that it is not always clear how to best incorporate *a priori* information about the local smoothness of the density into these estimators. Furthermore, these estimators usually break down in the tails where the data is sparse, and hence it is difficult to get good estimates of appropriate local bandwidths.

We propose a modification to the standard kernel estimator (1), first introduced in Rogers, Priebe, and Solka 1993, which uses a small number of bandwidths

instead of either extreme exemplified by equations (1) and (2).

Suppose we wish to have a small number of bandwidths where each bandwidth is associated with a region of the support of the density. To each bandwidth we associate a function which "filters" the data, in a sense to be described. Basically, the filter will define the extent to which each local bandwidth is to be used for any particular data point. We can then construct a kernel estimator which is a combination of the kernel estimators constructed using each bandwidth, with the data filtered by the filtering functions. To be specific, consider a set of functions $\{\rho_j\}_{j=1}^{m}$ where $0 \le \rho_j(x) \le 1$ for all x,

$$\sum_{j=1}^{m} \rho_j(x) = 1 \tag{3}$$

for all x as will be seen below. The ρ functions can be interpreted as posterior probabilities and are used to incorporate prior information concerning local smoothness. We will refer to the ρ functions as filtering functions. Associate to each filtering function ρ_j a bandwidth h_ij such that

$$\begin{aligned} 0 &< h_j \\ h_j &\to 0 \\ nh_j &\to \infty \end{aligned} \tag{4}$$

as n->∞. The filtered kernel estimator (FKE) for the filter $\{\rho_j\}_{j=1}^{m}$ is

$$\hat{\alpha}(x) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\rho_j(x_i)}{h_j}K\left(\frac{x-x_i}{h_j}\right). \tag{5}$$

The filtered kernel estimator comes from the following: given a finite mixture

$$f(x) = \sum_{j=1}^{m} \pi_j f_j(x) \tag{6}$$

and data $\{x_i\}$ with unknown density $\alpha(x)$, the kernel estimator filtered by the mixture f is defined to be (6) where

$$\rho_j(x) = \frac{\pi_j f_j(x)}{f(x)}. \tag{7}$$

The idea is to use a value of h for each component of f which is in some sense optimal for that component under the overall mixture model (6) and thus vary the bandwidth according to the individual variances of the filtering mixture. It is appealing to make use of the posterior probability of component membership as the local contribution for a given bandwidth. In practice, one would fit a mixture to the data which one felt was a good representative of the local variance of the underlying distribution, then use the mixture to construct bandwidths and a filtered kernel estimator. This approach works well even when the data is not distributed as a given finite mixture, provided that the mixture captures enough of the local variance characteristics of the data. Unfortunately, as will be seen, the calculation of the bandwidths h_j is not as simple as that for the standard kernel estimator (SKE), and requires the minimization of a function whose solutions are not known in closed form.

The $\rho_j(x_i)$ term in equation (5) weights the contribution of the kernel centered at x_i by its posterior component membership. This is shown in Figure 1, where a two component mixture density, an illustrative selection of kernels weighted by the

ρ_j functions, and the ρ_j functions themselves are shown.

An alternative to this formualtion can be obtained by considering a mixture of kernel estimators in which the posterior probabilities $\rho_j(x)$ at the point being estimated play the role of the mixing coefficients. This second approach allows the incorporation of information about the support of the density. This estimator is

$$\tilde{\alpha}(x) = \sum_{j=1}^{m} \rho_j(x) \left(\frac{1}{nh_j} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_j} \right) \right), \tag{8}$$

which can be rewritten in the form of the filtered kernel estimator as

$$\tilde{\alpha}(x) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\rho_j(x)}{h_j} K\left(\frac{x - x_i}{h_j} \right). \tag{9}$$

We incorporate information of the support of α by the condition that $\rho_j(x) = 0$ where α is known to vanish. We must have

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{h_j} \int \rho_j(x) K\left(\frac{x - x_i}{h_j} \right) dx = 1 \tag{10}$$

in order to guarantee that the estimate is a density. Since the proportions are not fixed, but are functions of x, we have a potentially different mixture for each x, which allows the incorporation of local scaling of the estimator.

We draw explicit attention to the dichotomous views demonstrated in (5) and (9). The estimator in (5) attributes the posterior component membership to the data point at which we center the kernel while (9) focuses on the component membership at the point at which we are computing the functional estimate. This type of dichotomy is inherent in the standard kernel estimator but in this case leads to dis-

tinct estimators.

We will assume throughout the formulation (5) for the filtered kernel estimator, unless otherwise noted. However, (9) might be of interest for those situations where the density is known to vanish, for instance for densities of known support. A combination of the two estimators (5) and (9) may in some situations be desired, but we not be pursued here.

Although we are concerned here with univariate densities, the filtered kernel estimator $\hat{\alpha}(x)$ has an interesting extension to multivariate densities. Assume that the kernel is a normal density and that the mixture (6) is a mixture of normals. For each local bandwidth h_j, we associate both the posterior probabilities from the mixture (the filtering function) and the covariance of the j^{th} component Σ_j. Then for each j, K is replaced with K_j, which is a normal with covariance Σ_j. Thus we can take into account local structure as represented by the mixture approximation to the density.

As always, we do not get something for nothing, and the filtered kernel estimator is no exception. Although the asymptotics make almost no restrictions on the choice of the filters and bandwidths, for finite samples these can be critical. In order for the filtered kernel estimator to provide any improvement over a single bandwidth kernel estimator (or anything else) we require filtering functions and local bandwidths which are appropriate for the density to be estimated. As will be shown in the examples below, this method works very well for densities that are approximated reasonably well by a mixture model, provided one has a good method for estimating the mixture model.

2. ASYMPTOTICS

Assume the conditions on the h_j's in eqn (4). Assume further that K(t) is a bounded density with zero mean and finite second moment k_2, that is,

$$\begin{aligned} \int tK(t)\,dt &= 0 \\ \int t^2 K(t)\,dt &= k_2 < \infty \end{aligned} \tag{11}$$

Thm 1: Under the above conditions, $\hat{\alpha}(x)$ and $\tilde{\alpha}(x)$ are weakly consistent.

pf:

Note that the theorem follows immediately from the consistency of the standard kernel estimator for the estimator $\tilde{\alpha}$. To show weak consistency for $\hat{\alpha}$ we need to show that both the bias and variance go to zero as n goes to infinity.

$$bias(\hat{\alpha}) = E\hat{\alpha} - \alpha \tag{12}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} E\left[\frac{\rho_j(x_i)}{h_j}K\left(\frac{x-x_i}{h_j}\right)\right] - \alpha(x)$$

$$= \sum_{j=1}^{m}\int\left[\frac{\rho_j(y)}{h_j}K\left(\frac{x-y}{h_j}\right)\right]\alpha(y)\,dy - \alpha(x)$$

$$\rightarrow \sum_{j=1}^{m}\rho_j(y)\,\alpha(y) - \alpha(y) = 0 \tag{13}$$

by Bochner's Lemma (Tapia and Thompson, 1978), since $h_j \to 0$. Similarly,

$$Var(\hat{\alpha}(x)) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(\sum_{j=1}^{m}\frac{\rho_j(x_i)}{h_j}K\left(\frac{x-x_i}{h_j}\right)\right)$$

$$\leq \frac{1}{n}E\left(\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{\rho_j(y)}{h_j}\frac{\rho_k(y)}{h_k}K\left(\frac{x-y}{h_j}\right)K\left(\frac{x-y}{h_k}\right)\right)$$

$$= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\int\frac{\rho_j(y)}{h_j}\frac{\rho_k(y)}{h_k}K\left(\frac{x-y}{h_j}\right)K\left(\frac{x-y}{h_k}\right)\alpha(y)\,dy$$

$$\leq \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{h_j h_k}\int K\left(\frac{x-y}{h_j}\right)K\left(\frac{x-y}{h_k}\right)\alpha(y)\,dy\ .$$

With a little manipulation we have

$$\leq \frac{sup(K(t))}{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{h_j h_k}\int K\left(\frac{x-y}{h_k}\right)\alpha(y)\,dy$$

$$\to \frac{sup(K(t))}{n}\sum_{j=1}^{m}\frac{1}{h_j}\ \alpha(x)\,k_2 \to 0 \qquad (14)$$

since $nh_j \dashrightarrow \infty$ for all j.

Thm 2: Under the same conditions as theorem 1 and assuming the existence of second derivatives of α and ρ_j, and that the second derivative of α is in L_2, the filtered kernel estimator $\hat{\alpha}(x)$ is L_2 consistent.

pf: Recall that the mean integrated squared error (MISE) can be written as

$$MISE(\hat{\alpha}) = \int bias^2(\hat{\alpha}) + Var(\hat{\alpha}). \tag{15}$$

So, we have

$$bias(\hat{\alpha}) = \sum_{j=1}^{m} \int \left[\frac{1}{h_j} K\left(\frac{x-y}{h_j} \right) \rho_j(y)\, \alpha(y) \right] dy - \alpha(x)$$

$$= \sum_{j=1}^{m} \int [K(t)\, \rho_j(x - h_j t)\, \alpha(x - h_j t)]\, dt - \alpha(x)$$

$$\approx \sum_{j=1}^{m} \int \left[K(t) \left\{ \alpha(x)\rho_j(x) - t h_j \frac{d}{dx}(\alpha(x)\rho_j(x)) + \frac{t^2 h^2}{2} \frac{d^2}{dx^2}(\alpha(x)\rho_j(x)) \right\} \right] dt - \alpha(x)$$

$$= \frac{k_2}{2} \sum_{j=1}^{m} h_j^2 \frac{d^2}{dx^2}(\alpha(x)\rho_j(x)), \tag{16}$$

and so

$$\int bias^2(\hat{\alpha}) \approx \frac{k_2^2}{4}\sum_{j=1}^{m}\sum_{k=1}^{m} h_j^2 h_k^2 \int \frac{d^2}{dx^2}(\alpha(x)\rho_j(x))\frac{d^2}{dx^2}(\alpha(x)\rho_k(x))\,dx. \quad (17)$$

Choosing $h_j=h_k=n^{-1/2}$ yields an of order $O(n^{-2})$.

In the case of the variance we have

$$Var(\hat{\alpha}) = \frac{1}{n}Var\left(\sum_{j=1}^{m}\frac{\rho_j(y)}{h_j}K\left(\frac{x-y}{h_j}\right)\right)$$

$$= \frac{1}{n}E\left[\sum_{j=1}^{m}\frac{\rho_j(y)}{h_j}K\left(\frac{x-y}{h_j}\right)\right]^2 - \frac{1}{n}\left\{E\left[\sum_{j=1}^{m}\frac{\rho_j(y)}{h_j}K\left(\frac{x-y}{h_j}\right)\right]\right\}^2$$

and so

$$Var(\hat{\alpha}) = \frac{1}{n}\int\left(\sum_{j=1}^{m}\frac{\rho_j(y)}{h_j}K\left(\frac{x-y}{h_j}\right)\right)^2 f(y)\,dy + O(n^{-1}) \quad (18)$$

$$\approx \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{1}{h_j}\frac{1}{h_k}\int K\left(\frac{x-y}{h_j}\right)K\left(\frac{x-y}{h_k}\right)\rho_j(y)\,\rho_k(y)\,\alpha(y)\,dy\ . \quad (19)$$

Letting

$$g(h_j, h_k) = \int K\left(\frac{t}{h_j}\right)K\left(\frac{t}{h_k}\right)dt, \quad (20)$$

we have

$$\int Var(\hat{\alpha}) \approx \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{m}\frac{g(h_j, h_k)}{h_j h_k}\int \rho_j(y)\,\rho_k(y)\,\alpha(y)\,dy\,. \quad (21)$$

Finally, assume without loss of generality that h_j is less than or equal to h_k, then

$$g(h_j, h_k) = \int K\left(\frac{t}{h_j}\right) K\left(\frac{t}{h_k}\right) dt$$

$$= h_j \int K(u)\, K\left(\frac{h_j}{h_k} u\right) du$$

$$\leq h_j \sup(K(t)) \int K\left(\frac{h_j}{h_k} u\right) du$$

$$= h_k \sup(K(t))$$

and so

$$g(h_j, h_k) \leq \min(h_j, h_k) \sup(K(t)), \tag{22}$$

so the integrated variance is of order $(n \min(h_k))^{-1}$ which can be made $O(n^{-1/2})$. Combining equations (17) and (21) we have

$$MISE \approx \left(\begin{array}{l} \frac{k_2^2}{4} \sum_{j=1}^{m} \sum_{k=1}^{m} h_j^2 h_k^2 \int \frac{d^2}{dx^2}(\alpha(x)\rho_j(x)) \frac{d^2}{dx^2}(\alpha(x)\rho_k(x))\, dx + \\ \qquad \frac{1}{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{g(h_j, h_k)}{h_j h_k} \int \rho_j(y)\, \rho_k(y)\, \alpha(y)\, dy \end{array} \right. \tag{23}$$

Thus, MISE $\rightarrow$ 0, at a rate no worse than $O(n^{-1/2})$. As in the case of the standard kernel estimator, the optimal rate is $O(n^{-4/5})$.

Thm 3: With the same conditions as Theorem 2, $\tilde{\alpha}$ is L_2 consistent.

pf:

Following the calculations in theorem 2 and noting that the ρ_j are bounded above by 1, we have

$$\int bias^2(\tilde{\alpha}) \approx \frac{k_2^2}{4} \sum_{j=1}^{m} \sum_{k=1}^{m} h_j^2 h_k^2 \int \rho_j(x)\, \rho_k(x)\, (\alpha''(x))^2 dx, \tag{24}$$

which is of order $O(n^{-2})$.

For the variance, we have

$$\int Var(\tilde{\alpha}) \leq \frac{1}{n} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{g(h_j, h_k)}{h_j h_k}, \tag{25}$$

which is once again of order $(n\ min(h_k))^{-1}$. Thus we have MISE $\rightarrow 0$ at a rate of $O(n^{-1/2})$, with optimal rate $O(n^{-4/5})$.

For the rest of this paper we will consider the estimator (5), which will be referred to as the FKE. Note that given any filter, for the optimal choice of the h_j we have $MISE_{FKE} \leq MISE_{SKE}$. This is a trivial consequence of the fact that the FKE subsumes the SKE, when we take all the h_j's to be equal.

3. SPECIAL CASE: NORMAL KERNELS

We now assume that the kernel K is the standard normal. In this case we can com-

pute g() and obtain

$$g(h_j, h_k) = \frac{1}{\sqrt{2\pi}} \frac{h_j h_k}{\sqrt{h_j^2 + h_k^2}}. \tag{26}$$

In keeping with the ideas discussed in the introduction, we assume in this section that α is a mixture of normals, and that the filtering functions are generated by the same mixture. Equation (23) then becomes, using the notation of equation (6)

$$MISE \approx \begin{cases} \frac{k_2^2}{4} \sum_{j=1}^{m} \sum_{k=1}^{m} \pi_j \pi_k h_j^2 h_k^2 \int f''_j(x) f''_k(x)\, dx + \\ \frac{1}{n\sqrt{2\pi}} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{\pi_j \pi_k}{\sqrt{h_j^2 + h_k^2}} \int \frac{f_j(y) f_k(y)}{\alpha(y)} dy \end{cases} \tag{27}$$

At this point we introduce some notation.

$$A_{jk} = \pi_j \pi_k \int f''_j(x) f''_k(x)\, dx\ , \tag{28}$$

$$B_{jk} = \pi_j \pi_k \int \frac{f_j(x) f_k(x)}{\alpha(y)} dx. \tag{29}$$

This gives the equation

$$MISE \approx \frac{k_2^2}{4} \sum_{j=1}^{m} \sum_{k=1}^{m} A_{jk} h_j^2 h_k^2 + \frac{1}{n\sqrt{2\pi}} \sum_{j=1}^{m} \sum_{k=1}^{m} \frac{B_{jk}}{\sqrt{h_j^2 + h_k^2}}. \tag{30}$$

Taking the partial with respect to h_r we have

$$\frac{\partial}{\partial h_r} MISE = k_2^2 A_{rr} h_r^3 + \frac{1}{2} k_2^2 h_r \sum_{k \neq r} A_{kr} h_k^2 - \frac{B_{rr}}{2n\sqrt{\pi} h_r^2} - \frac{h_r}{n\sqrt{2\pi}} \sum_{k \neq r} \frac{B_{kr}}{\left(\sqrt{h_r^2 + h_k^2}\right)^3} . \quad (31)$$

Equations (30) and (31) are used in the next section to compute the optimal bandwidths and MISE of a number of examples. Note that this must be done numerically, since we do not have a closed form solution to the problem of minimizing (30).

In practice the true underlying mixture is not known, and in fact the data may not come from a mixture at all. In this case it may not be clear how to apply the above formulation and calculate the desired local bandwidths, not the least because the A_{jk} and B_{jk} require α to be a known mixture. We propose the following approach to this problem: we first approximate the unknown density as a mixture, then minimize (30) to calculate the bandwidths under the assumtion that the filtering density is the true density. Thus we use the optimal values for h_j under the assumption that the filtering mixture is correct. This is analogous to using a reference density such as a normal to compute the bandwidth for the standard kernel estimator. As in the case of the standard kernel estimator, our estimate will only be optimal if the filtering mixture is indeed correct, but it will be a useful estimate as long as the data is close to the filtering mixture.

4. EXAMPLES

We compare the MISE of the FKE with the standard kernel estimator with h chosen optimally. When simulations are performed, the bandwidths are chosen by numerically minimizing (30). Following Wand, Marron, and Ruppert 1991, we compute the efficiency of the estimator as $MISE_{FKE}/MISE_{SKE}$ so small values of the efficiency correspond to better estimates with the FKE.

Case 1: Two means.

Let

$$\alpha(x) = \frac{1}{2}N(0,1) + \frac{1}{2}N(m,1) \tag{32}$$

It is easy to see that in this case the optimal bandwidth choice for the FKE requires $h_1 = h_2 = h_{SKE}^{opt}$, and so $MISE_{FKE} = MISE_{SKE}$ and the efficiency is 1. This is intuitively what should happen, since the FKE is designed to incorporate differences in variance of the underlying mixture components, and so it will give no improvement in cases where the components differ only in mean.

Case 2: Two variances.

Let

$$\alpha(x) = \frac{1}{2}N(0,1) + \frac{1}{2}N(0,v)\ , \tag{33}$$

with $.1 \leq v \leq 10$. Figure 2a shows the efficiency as a function of the variance. Note that for $v \neq 1$, the FKE improves on the SKE, as one would expect. Figure 2b shows the two bandwidths used in the FKE. The bandwidth associated with the second mixture term, the term for which we vary v, dramatically changes accord-

ing to v.

This is essentially the case that the FKE was designed to address. We have a density which is a mixture of two normals with unequal variances. As the variance of the second term is moved away from the variance of the first term, the standard kernel's single bandwidth becomes less and less appropriate for the resulting density. The filtered kernel estimator allows us to take the two variances into account in our estimator, thus improving the estimate when the variances are significantly different.

Case 3: Outlier model

Let

$$\alpha(x) = pN(0, 1) + (1-p)N(0, 100), \tag{34}$$

with $.01 \leq p \leq .99$. Figure 3shows the efficiency for this model as a function of p. Again we see that the FKE improves over the standard kernel provided $0 < p < 1$. Clearly the estimators are equal when p=0 or p=1.

Case 4: Marron and Wand Densities

Marron and Wand 1992 list 15 normal mixture densities showing some of the wide range of variations that are obtainable with simple mixtures. Table 1 shows the efficiency of the FKE for these densities. Note that the performance of the FKE depends on the amount of local variability of the mixture, as would be expected.

The above examples dealt with the theoretical properties of the FKE, where the filter is assumed to be equal to the underlying density. In practice this is not possible, and in fact if the underlying density is known any attempts at estimation are obviously unnecessary. In the next two examples we consider the case where the underlying density is not known, and in fact in at least one (case 5) it is known not to be a mixture of normals at all. In these cases we first fit a mixture to the data to obtain a reasonable filter. Then we compute the h_j under the assumption that the filter is equal to the density. In practice, as will be seen below, this provides a good estimator provided the filtering mixture captures most of the underlying variability of the data.

Case 5: Lognormal.

100 data points were drawn from a lognormal and a two component mixture was fit to the data using the EM method (see, e.g., Titterington, Smith, and Makov 1985). The bandwidths for the filtered kernel estimator were chosen assuming the filter to be equal to the true density. Thus we first construct the mixture estimate and then use the bandwidths that would be optimal for that mixture density, in much the same way that one might use a normal density as a reference estimate for the standard kernel estimator.

Figure 4a shows the density estimates for the standard kernel estimator and the FKE. The bandwidths for the FKE were $h_1 = .4$ and $h_2 = 2.2$. The bandwidth for the standard kernel estimator was chosen by hand to get a reasonable fit to the true density. The plot shown uses a value of .2 for h, which is about five times the

"optimal" value for a lognormal using 100 data points. Figure 4b compares the density estimates of the FKE and the filtering mixture.

Case 6: Suicide Data.

Silverman 1986 uses a data set of lengths of treatment spells in days of control patients in a suicide study to illustrate the kernel estimator. We fit two normals to the data using the EM algorithm and use this as the filter for the FKE, as was done with the lognormal example. We compare the estimator with the two kernel estimators Silverman uses, bandwidths = 20 and 60 in figure 5a. The FKE uses the bandwidths $h_1 = 19.17$ at the mode and $h_2 = 127.17$ in the tail. It is noteworthy that these bandwidths correspond to Silverman's choices of h=20 to get pleasing results in the mode at the expense of tail smoothness and double Silverman's h=60 chosen to yield a smooth tail. The FKE is compared with the mixture approximation in Figure 5b. Note that the FKE allows a good fit to the mode while maintaining the smoothness of the tail. The mode smoothness can be varied by varying the appropriate bandwidth (h_1) without having much effect on the fit to the tail, as can be seen by the plot of the filtering functions in Figure 5c. This figure makes clear the local character of the bandwidths in the FKE. This is also illustrated in Figure 5d, where the bandwidth associated with the mode is reduced, making the mode more pronounced and rough without effecting the tail smoothness. It is this ability that makes the FKE a very interesting, and we feel useful, estimator.

5. CONCLUSIONS

The filtered kernel estimator is superior in performance to the standard kernel estimator, provided appropriate filter functions and bandwidths can be chosen. In section 2 it was shown that any filter functions will give asymptotic performance no worse than the standard single bandwidth kernel estimator.

It would seem at first that the added trouble of selecting filtering functions and bandwidths would make the estimator difficult to use in practice. However the idea of using a finite mixture fit to the data to construct the filters is one which appears to work well in a variety of situations, even those for which the data is not drawn from a finite mixture. The ability to take local structure into account is a powerful one which will allow much better estimates in those situations where there is reason to believe the local structure is justified.

It should be noted that bad filters produce bad FKE's. This is not unreasonable, however it does mean that care must be used in the choice of the filtering mixture. Just as the standard kernel estimator produces errors when the bandwidth is taken to be too large or too small, mixtures which have terms which are not supported by the data will produce local errors in the FKE estimate. This local character of the estimator gives some protection, since the effect of the error is reduced outside the region in which the corresponding ρ dominates. This is in contrast to the single kernel estimator where the choice of the bandwidth has a global effect.

We have focused on the univariate case in this work, but other extensions are possible. The multivariate version of the FKE has much promise, and will be addressed in the future. The ability to effectively tune the kernels to the local structure of the data will be a powerful and useful tool for multivariate density estimation. It is believed that this ability to define the structure locally will be of use in

exploritory data analysis and in discriminant analysis.

REFERENCES

Marron, J.S. and Wand, M.P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712-736.

Parzen, E. (1962), "On Estimation of a Probability Density Function and Mode," *Ann. Math. Statist.*, 33, 1065-1076.

Rogers, G.W., Priebe, C.E., and Solka, J.L. (1993), "Filtered Kernel Probabilistic Neural Network," *SPIE Vol. 1962 Adaptive and Learning Systems II*, 242-252.

Rosenblatt, M. (1956), "Remarks on some Nonparametric Estimates of a Density Function," *Ann. Math. Statist.*, 27, 832-835.

Scott, D.W. (1992), *Multivariate Density Estimation*, New York: John Wiley.

Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman and Hall.

Tapia, R.A. and Thompson, J.R. (1978), *Nonparametric Probability Density Estimation*, Baltimore: The Johns Hopkins University Press.

Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985), *Statistical Analysis of Finite Mixture Distributions*, New York: John Wiley.

Wand, M.P., Marron, J.S., and Ruppert, D. (1991), "Transformation in Density Estimation," *Journal of the American Statical Association*, 86, 343-361.

Efficiency results for the FKE for the 15 normal mixture densities from Marron and Wand 1992. These densities show some of the wide range of variations that are obtainable with simple mixtures. The FKE is superior to that of the standard kernel estimator when there is variability of the mixture's local variance structure.

Table 1:

Density	Efficiency
Gaussian	1
Skewed Unimodal	.60
Strongly Skewed	.38
Kurtotic Unimodal	.44
Outlier	.91
Bimodal	1
Separated Bimodal	1
Skewed Bimodal	.69
Trimodal	.90
Claw	.63
Double Claw	.13
Asymmetric Claw	.30
Asymmetric Double Claw	.15
Smooth Comb	.41
Discrete Comb	.5

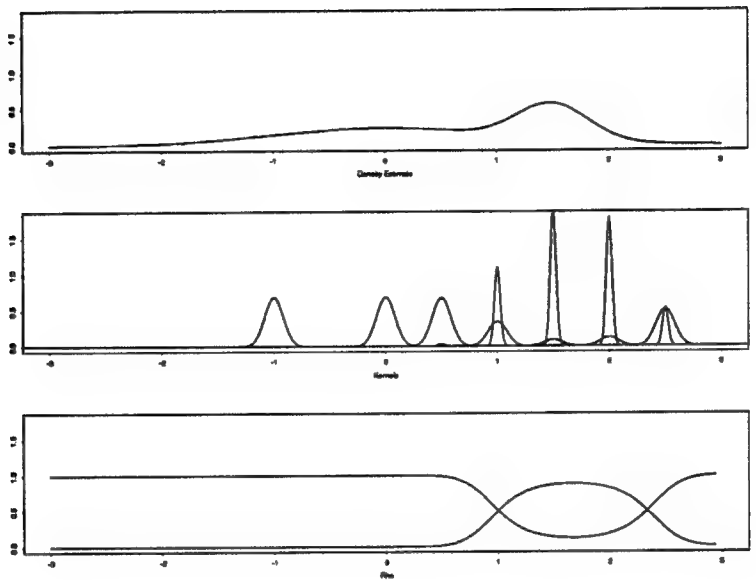


Figure 1

An example of the filtered kernel estimator applied to a two component mixture. The mixture probability density function, weighted kernels, and posterior ρ functions, are shown.

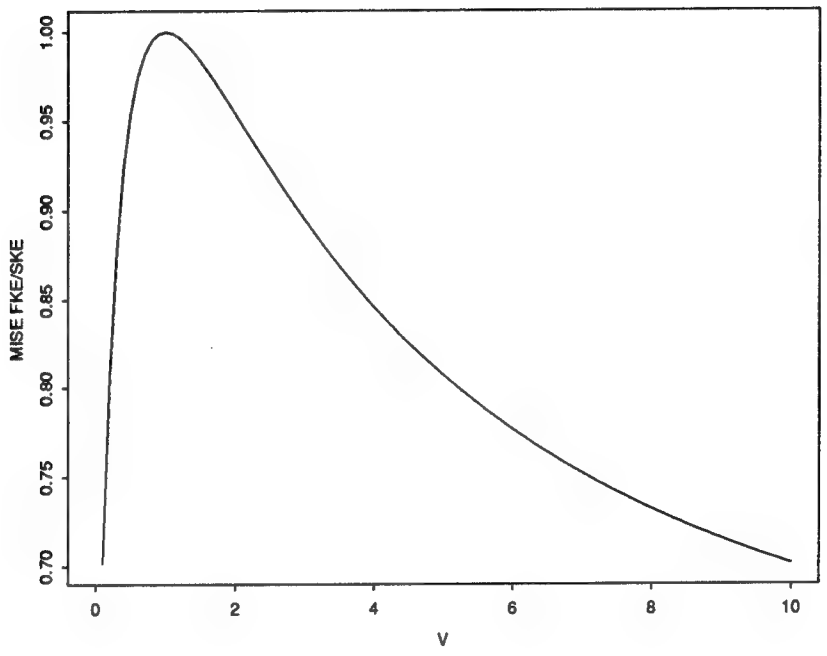


Figure 2a

Efficiency as a function of variance for case 2: $\alpha(x) = \frac{1}{2}N(0,1) + \frac{1}{2}N(0,v)$. As v deviates from 1 the standard kernel's single bandwidth becomes less and less appropriate for the resulting density. The filtered kernel estimator allows us to take the two variances into account in our estimator, thus improving the estimate when the variances are significantly different.

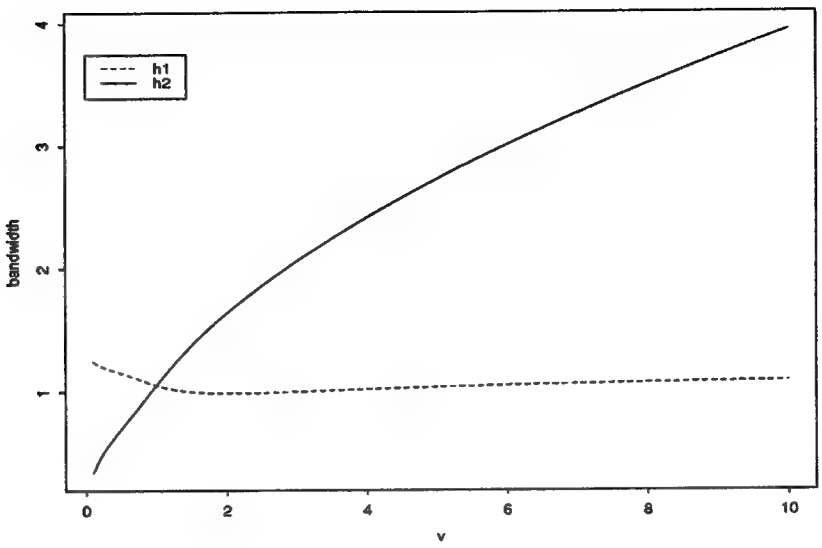


Figure 2b

The two bandwidths used in the FKE. The bandwidth associated with the second mixture term (solid line), the term for which we vary v, dramatically changes with v allowing the FKE to model the local variance structure of the underlying density.

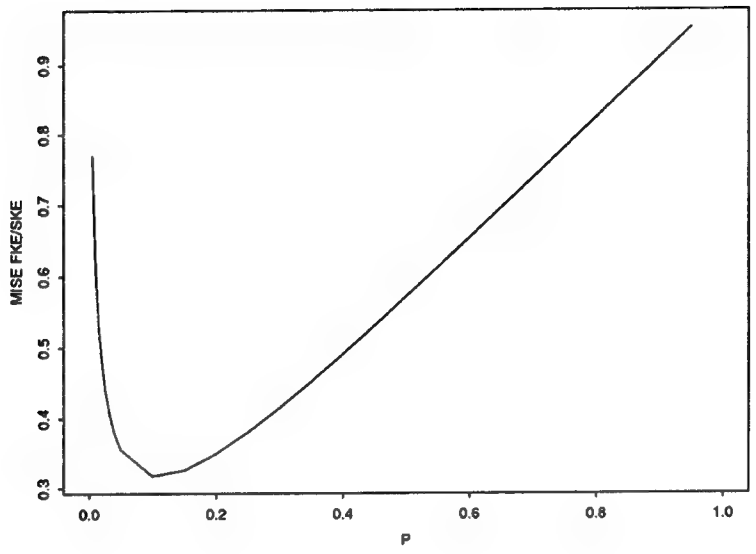


Figure 3

Efficiency as a function of p for the outlier model (case 3): $\alpha(x) = pN(0,1) + (1-p)N(0,100)$. As p runs from .01 to .99 the FKE improves over the standard kernel estimator because the underlying density has nonconstant local variance structure.

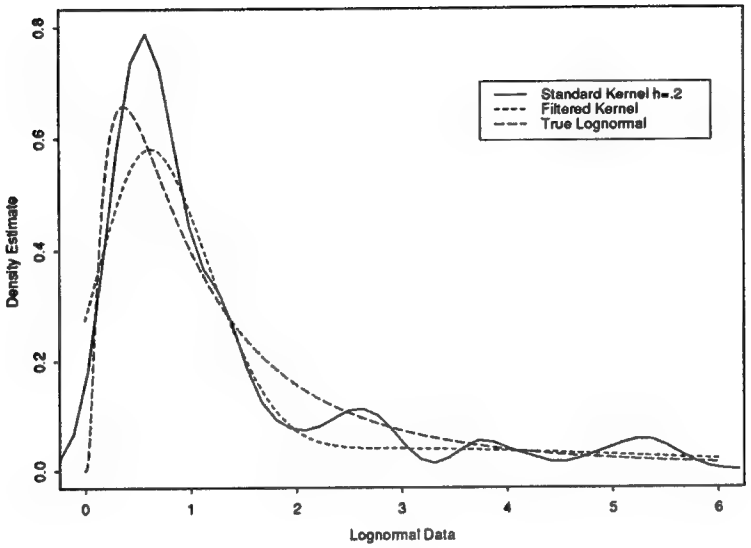


Figure 4a

Density estimates for the standard kernel estimator and the FKE, along with the true lognormal density from which 100 sample observations were drawn (case 5). The bandwidths for the FKE are $h_1 = .4$ and $h_2 = 2.2$. The bandwidth for the standard kernel estimator (h=.2) was chosen to get a reasonable fit to the true density.

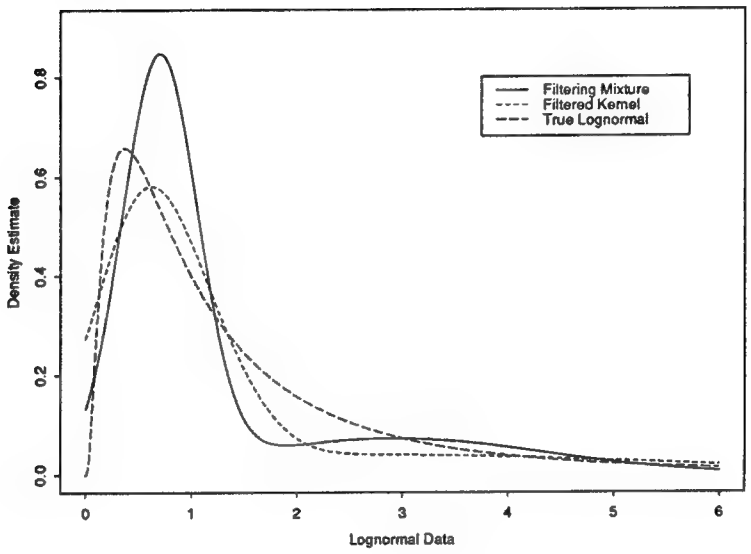


Figure 4b

The FKE, the filtering mixture, and the true lognormal density from case 5.

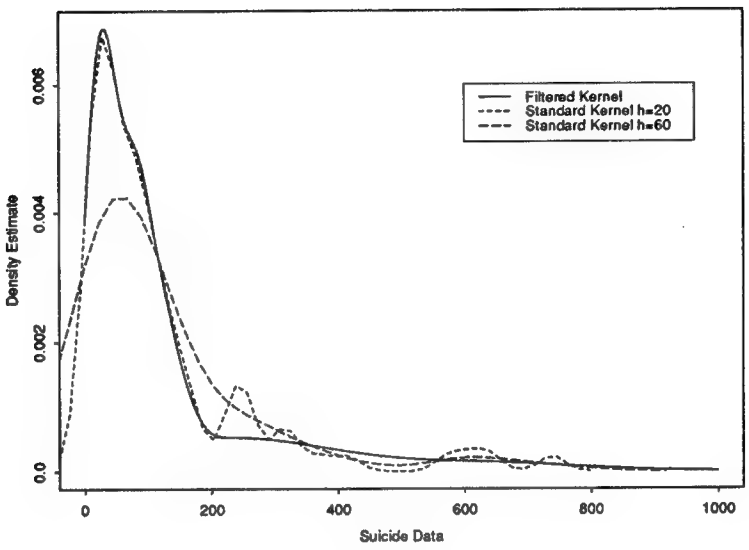


Figure 5a

Case 6: suicide data from Silverman 1986. A mixture of two normals is used as the filter for the FKE. The FKE is compared with standard kernel estimators using bandwidths of 20 and 60. The FKE uses the bandwidths $h_1 = 19.17$ at the mode and $h_2 = 127.17$ in the tail and combines the features of the two SKEs - detail in the mode and smoothness in the tail.

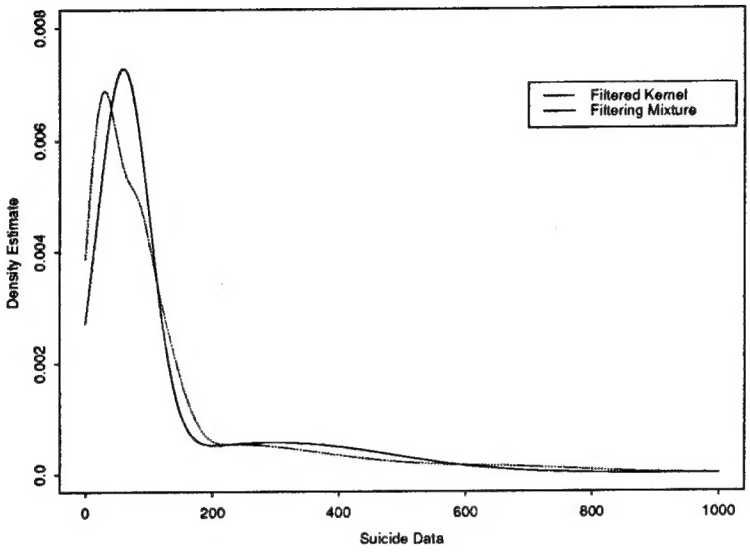


Figure 5b

Comparison of the FKE from Figure 5a and its associated mixture approximation.

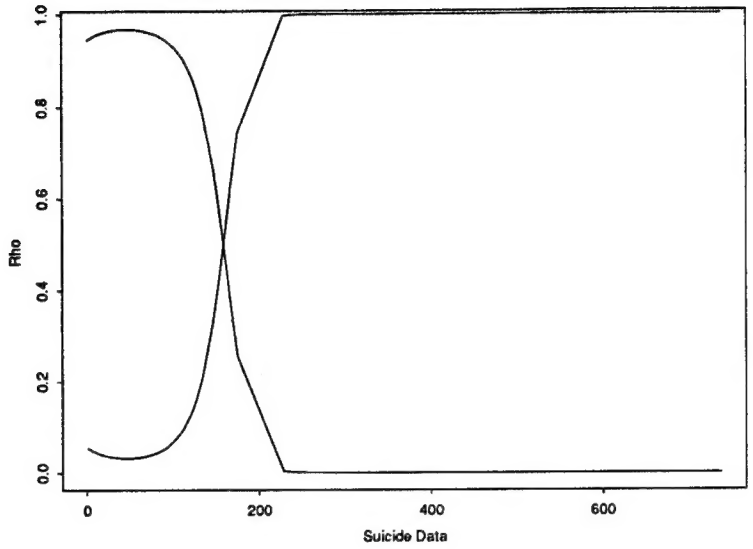


Figure 5c

The filtering functions ρ defined by the mixture estimate shown in Figure 5b. These posterior functions dictate the local character of the bandwidths in the FKE, and indicate that the local smoothness can be varied by varying the appropriate bandwidth without having much effect on the fit in other regions.

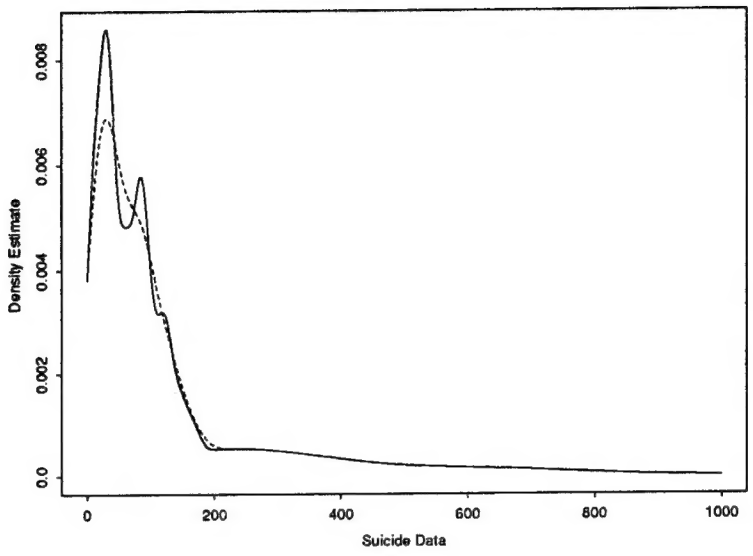


Figure 5d

An example of local selective tuning. For the FKE shown in Figure 5a (dashed line) the bandwidth associated with the mode has been reduced, making the mode more pronounced and rough without effecting the tail smoothness. This is due to the effect of the filtering functions shown in Figure 5c.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October, 1994	Technical Report

4. TITLE AND SUBTITLE
Filtered Kernel Density Estimation

5. FUNDING NUMBERS
DAAH04-94-G-0267

6. AUTHOR(S)
David J. Marchette, Carey E. Priebe,
George W. Rogers and Jefferey L. Solka

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for Computational Statistics
George Mason University
Fairfax, VA 22030

8. PERFORMING ORGANIZATION REPORT NUMBER
TR no. 104

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARO 32850.6-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A modification of the kernel estimator for density estimation is proposed which allows the incorporation of local information about the smoothness of the density. The estimator uses a small set of local bandwidths rather than a single global one as in the standard kernel estimator. It uses a set of filtering functions which determine the extent of influence of the local bandwidths. Various versions of the idea are discussed. The estimator is shown to be consistent and is illustrated by comparison to the single bandwidth kernel estimator for the case in which the filter functions are derived from finite mixture models.

14. SUBJECT TERMS	15. NUMBER OF PAGES
local bandwidth, finite mixtures, EM algorithm MISE	34
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL